NIST 2010 Open Handwriting Recognition and Translation Evaluation Plan

version 2.8

1 Introduction

The National Institute of Standards and Technology (NIST) Open Handwriting Recognition and Translation (OpenHaRT) evaluation is a public evaluation of image-to-text transcription and translation, similar to the tasks evaluated by NIST for the DARPA Multilingual Automatic Document Classification Analysis and Translation (MADCAT)[1] Program [1]. The 2010 evaluation will focus on recognition and translation of images containing primary Arabic handwritten script as we explore the tight coupling between optical character recognition (OCR) and machine translation (MT) technologies. We seek to break new ground in the area of document image recognition and translation toward the goal of document understanding. The objective of this evaluation is to build the critical mass required to solve challenges posed in these areas.

Participation in the evaluation is open to all who find the tasks of interest. There is no fee to participate. However, participants are required to attend the post-evaluation workshop[2] to present their systems. While the evaluation is open to all who wish to participate, workshop attendance is limited to evaluation participants, data providers, (potential) evaluation sponsors and interested government personnel.

To participate in the evaluation, interested parties must register by completing the registration form[3] available at the NIST OpenHaRT web site.

2 Evaluation Tasks

Technologies that OpenHaRT seeks to advance are multidisciplinary and include OCR and MT. The goal of OpenHaRT is to assess system performance and to understand the strengths and weaknesses of particular algorithmic approaches. It is planned that future evaluations will build on these technologies to include more complex tasks that are required to achieve document understanding capabilities.

Segmentation plays a vital role in deconstructing the document images. Translation and recognition tasks are paired with segmentation conditions to explore the relationship between system performance and the system's ability to segment the data. Segmentation is represented as a series of polygon coordinates indicating the locations of the text segments within the image. The two segmentation conditions in OpenHaRT'10 are referred to as *word segmentation* and *line segmentation.*

- **Word segmentation** is created manually. Human annotators mark the word boundaries using the GEDI [2] tool. The input to GEDI is a document image.

- **Line segmentation** is the primary segmentation condition. It is defined as a bounding box that surrounds a line of text and is derived algorithmically from the word segmentations by creating polygons that minimize the amount of text overlap between the lines.

There are three separate tasks defined for inclusion in OpenHaRT 2010. Each task was designed to measure components within the overall system. Tasks are described below and summarized in the appendix (Table **4**). Participation in any one, two or all three tasks is encouraged.

2.1 Document Image Translation

Document image translation measures the overall performance of the system in translating text in foreign language document images into accurate and fluent English. This task is offered for both word and line segmentation conditions. The system is given a document image and the appropriate level of segmentation information and is required to output the English translation. Refer to sections 6.2 and 6.3 for input and output format. Re-rendering the image is not required and is not a focus for OpenHaRT at this time.

2.2 Document Image Recognition

Document image recognition measures the text recognition component (OCR) of the system transcribing the handwritten text in the document

[1] The NIST OpenHaRT evaluation is closely related to the DARPA MADCAT evaluation. Thus there will be many references to "MADCAT" throughout this document.

[2] The workshop will be held in the Baltimore/Washington D.C. area. There is a small workshop registration fee which does not include travel and accommodation.

[3] http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2010_RegistrationForm.pdf

image. This task is offered for both word and line segmentation conditions.

2.3 Document Text Translation

Document text translation measures the translation component (MT) of the system given a manually produced transcription of the text in the document image. Image segmentation condition is not applicable to this task.

3 Data Resources

The data being used for OpenHaRT 2010 was created by the Linguistic Data Consortium (LDC) and has been used in previous MADCAT evaluations. This data was created in a controlled environment where known scribes copied Arabic source texts that were previously used in the DARPA GALE [2] program.

A set of LDC-collected corpora will be provided for system development. To receive this data resource, participants must register for the 2010 OpenHaRT evaluation and sign the LDC license agreement[4] acknowledging the terms governing the use and rights to the data.

Participants are free to use additional non-LDC collected data for system development. However, participants should take the necessary precautions to exclude newswire and web data that was originally published within the evaluation epoch **June 1-30, 2007.** We acknowledge that this exclusion may not always be practical. Participants are required to document data used for system development in their system description. See 8.2 for more information about system descriptions.

3.1 Document Sources

The source data comes from a variety of Arabic newswire publications, web blogs and online discussion forums. Newswire (NW) represent formal or structured text while and web text (WB) represents informal or unstructured text.

3.2 Document Image Creation

The source text was originally in electronic format. A corresponding document image was created by instructing literate native Arabic writers to produce handwritten copies of chosen passages using various writing conditions. Each passage was copied by at least two scribes. The handwritten copies were then scanned at 600 dpi to create the document images in TIFF format. Table 1 lists the target distribution of the various writing factors, and Table 2 lists statistics for the datasets.

Table 1: Target distribution of various writing factors

Writing Instrument	Writing Surface	Writing Speed
90% ballpoint pen	75% unlined whitepaper	90% normal
10% pencil	25% lined paper	5% fast
		5% careful

Table 2: Data profile for OpenHaRT datasets. "P1" and "P2" refer to phase1 and phase2 MADCAT evaluations.

	Training Set	Dev Set	Eval Set
Source	MADCAT P1/P2 training & devtests	MADCAT P1 pilot eval set	MADCAT P2 eval set
Genre	NW & WB		
Num. of passages	~6000	~100	~100
Arabic tokens per passage	125	125	125
Number of scribes per passage	Max 5	2	3
Scribe distribution	50% exposed, 50% unexposed		
Total num of scribes	100	24	24

4 Evaluation Metrics

This section describes the metrics used to score each of the three evaluation tasks.

4.1 TER

The system performance on the **document image translation and document text translation** tasks is measured automatically using the official evaluation metric Translation Error Rate (TER) [4]. TER is an edit distance metric which calculates the exact match distance between the system translation and the reference translation.

$$TER = \frac{(\#insertions + \#deletions + \#substitutions + \#shifts)}{\#reference\_translated\_words}$$

In addition, system performance will be measured using a set of alternative automatic metrics such as BLEU and METEOR.

4.2 WER

The system performance on the **document image transcription** task is measured automatically using the Word Error Rate (WER) [5] metric. WER is an

[4] http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2010_LDCLicenseAgreement.pdf

edit distance metric which calculates the errors (insertions, deletions, and substitutions) in the system transcription. (*TER and its use of shifts are not applicable to measuring transcription errors.*)

$$WER = \frac{(\#insertions + \#deletions + \#substitutions)}{\#reference\_transcribed\_words}$$

5 Scoring Package

NIST is developing a scoring package to facilitate the calculation of the OpenHaRT metrics. The package utilizes the software terp.v1 developed by UMD-BBN [4] as well as those developed internally at NIST [5]. The availability of the package will be announced on the OpenHaRT mailing list hart_list@nist.gov.

Normalization is to be performed on the system output prior to scoring. For the translation tasks, punctuations in the reference and system translations are tokenized. For the transcription task, if any diacritic information is present in the reference and system transcripts, it is removed.

Segments containing scribe errors (e.g., typos, word omissions) are to be included as-is for scoring. A stand-off annotation file will identify such segments allowing them to be analyzed separately.

All translation and transcription scoring preserves the casing information.

6 Data File Format

OpenHaRT data use an XML format that defines storage elements which capture the various annotation layers in a document image. The format is described in version (v6) of the MADCAT Format Specifications document[5] and is extendable to future planned evaluation tracks. All training, development, and evaluation data will adhere to this XML format. System output will be validated using the DTD version 1.1.1 before being scored.

6.1 Reference Data

Each reference file contains two main layers of information along with a pointer to an accompanying document image. The first layer contains the physical segmentation of the image. The second layer contains semantic information in the image. The reference files are identified with the extension ".madcat.xml".

[5] http://www.nist.gov/itl/iad/mig/upload/MADCATDataFormatSpec_V6-tgz.txt. This is a tar'd compressed file (tgz) but was renamed so it could be uploaded to the NIST webspace. Rename "-tgz.txt" to ".tgz" and uncompress, untar as normal.

For example: <FILENAME>.madcat.xml

6.2 Input Data

The input to the system under test consists of document images and their corresponding XML files identifying the segmentation of interest.

The XML input files are derived from the reference files. Depending on the task, certain information will be removed from the reference files to create the input files. For the document image translation and the document image transcription tasks, the translation and transcription information is removed. For the document text translation task, translation information is removed. If a task excludes some segmentation information, the corresponding segmentation sub-layer is also removed.

Table 5 in the appendix summarizes the information content in the input for each task-condition pairing. Note the filename extensions also indicate the information content in the input.

6.3 Output Data

The output from the system under test consists of input XML files with the missing information added by the system.

Depending on the task, certain information will be added to the input files to create the output files. For the document image translation and the document image transcription tasks, the OpenHaRT system is to add the missing translation and transcription information, respectively. For the document text translation task, the OpenHaRT system is to output the translation information.

The name of the output file is to follow a similar naming convention as its corresponding input file. Note the difference between input and output filenames is highlighted below in red. Refer to Table 5 in the appendix for the output filename convention.

For example:

Input: <FILENAME>.wordseg.madcat.xml

Output: <FILENAME>.wordseg.sys.madcat.xml

7 Evaluation Rules

The following rules must be observed when participating in the OpenHaRT evaluations:

- All tasks must be processed independently. That is, data files provided to complete other evaluation tasks may only be used for their designated task.

- Language model adaptation across pages is not allowed.

- Investigation of the evaluation data prior to submission of system output is not allowed. Both human and automatic probing is prohibited.

- To the extent possible, participants must exclude data that overlaps the evaluation epoch of June 1 – 30, 2007 from system development.

- Participation in the post-evaluation workshop is required. Each participating organization is to be represented by at least one technical individual who has the knowledge required to discuss system details (algorithmic approaches, data, issues ...) in the workshop's open forum.

8 Submission of Results

Participants may submit results for more than one system. If more than one system is submitted for a given task/condition pairing, one system must be declared the primary system at the time of submission. All other systems for the same task are to be considered contrastive[6] systems.

8.1 System Output Submission

Submission of results will be made via FTP. Submissions that fail DTD validation will be returned to participants for correction. Late and/or debugged submissions will be documented and scored but will not be compared to other on-time submissions in NIST's reports.

Participants are to follow the steps outlined below when submitting their results.

1) Create a directory where the system output will reside. The directory is to follow the format:

<TEAM>_<TASK>_<SYS>_<SUB_NUM>, where

TEAM: is the name of the participating team
TASK: is one of the following abbreviated task names:
- **DIT** – document image translation
- **DIR** – document image recognition
- **DTT** – document text translation

SYS: is the name of the system:
- **p** – primary system
- **c**N – contrastive system where N is an integer

SUB_NUM: is an integer indicating the submission number.

2) Place the system output in that directory
3) Tar and compress the directory
4) FTP the compressed tar file to jaguar.ncsl.nist.gov/openhart/incoming
5) Send an email to hart_poc@nist.gov to notify the submission was made

For example:

- mkdir NIST_DIT_c1_1
- cp *.{wordlline}sys.madcat.xml NIST_DIT_c1_1
- tar zcvf NIST_DIT_c1_1.tgz NIST_DIT_c1_1
- ftp jaguar.ncsl.nist.gov (anonymous login with email as password)
  - binary
  - cd hart/incoming
  - put NIST_DIT_c1_1.tgz
  - bye
- send an email to hart_poc@nist.gov

8.2 System Description Submission

In addition to the system output, sites are to include a system description describing the system(s) submitted for evaluation. The system description consists of, but is not limited to, the algorithm approaches employed, the training data used, and/or any other pertinent information. A template for the system description[7] can be obtained from the NIST OpenHaRT web site.

Submission of the system descriptions will be made via email. There should be only one system description per participating team with the name:

<TEAM>_sysdesc.<EXTENSION>, where

TEAM: is the name of the participating team (same as in 8.1)
EXTENSION: is file extension and must be one of the following formats:
- **txt** – ASCII text
- **doc** – Microsoft Word
- **pdf** – Adobe PDF

The system description is to be sent as an email attachment to hart_poc@nist.gov. Refer to section 10 for the system description due date.

[6] The term "contrastive" indicates the system serves as a contrast to the main or "primary" system. The contrast could be in the training data used or in the algorithmic approaches employed. A site's contrastive systems will only be compared against its primary system for the same task.

[7] http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2010_SystemDescription.txt

9 Publication of Results

NIST will release an official scoring report following the evaluation workshop. The report will be made public on the NIST website. Participants are free to publish and discuss their own results. However, participants must not publicly compare their results to that of other participants but can point to the NIST report for the results of the other participants. Participants must reference the NIST report when publishing their results.

10 Schedule

Table 3 lists important dates of the evaluation. Participating sites will receive training data when the registration form and data license agreement are completed.

Table 3: OpenHaRT'10 evaluation schedule

Event	Date
Evaluation epoch (development data cannot overlap this epoch)	June 1-30, 2007
Evaluation registration deadline	June 15, 2010
Evaluation period	July 19-Aug 4, 2010
DIT/DIR data distributed to participants	*July 19 (~9am Eastern)*
DIT/DIR results due to NIST	*July 26 (~9am Eastern)*
Submissions verification	*July 26-27*
DTT data distributed to participants	*July 28 (~9am Eastern)*
DTT results due to NIST	*Aug 4 (~5pm Eastern)*
Preliminary results released to participants	Aug 6, 2010
System description due to NIST	Aug 20, 2010
Post-evaluation workshop	Sept 16, 2010
Official results published	Oct 8, 2010

11 Glossary of Terms

The following terms are used throughout the document and are defined below for clarifications.

- **Document** – a naturally occurring unit of original source data of variable length
- **Passage** – a sub-section within a document chosen for evaluation
- **Manuscript** – a copy of a passage created by a scribe
- **Page** – one of the leaves in a manuscript created by a scribe; the basic unit of evaluation
- **Scribe** – a person who creates a handwritten copy of one or more passages

12 References

[1] J. Olive, "Multilingual Automatic Document Classification Analysis and Translation (MADCAT) SOL BAA 07-38 Proposer Information Pamplet", DARPA/IPTO, 2007.

[2] E. Zotkina, H. Suri, D. Doermann, "GEDI: Groundtruthing Environment for Document Images (Software)", http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=53.

[3] GALE_p3_evalplan-v1f.pdf at http://www.nist.gov/itl/iad/mig/upload

[4] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," *Proceedings of Association for Machine Translation in the Americas*, 2006.

[5] J. Fiscus, J. Ajot, N. Radde, and C. Laprun, "Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech", *Proceedings of LREC*, 2006.

Appendix

Table 4: Overview of OpenHaRT'10

Task	Primary Metric	Input	Output
Document Image Translation	TER	Arabic document image • with line segmentation • with word segmentation	Segmented English translation
Document Image Recognition	WER	Arabic document image • with line segmentation • with word segmentation	Segmented Arabic transcription
Document Text Translation	TER	Segmented manual transcription of Arabic document image	Segmented English translation

Table 5: Information content for the evaluation tasks

Task	Condition	Annotation Layer Removed	Input/Output File Extension
Document Image Translation	Word Segmentation	• transcription • translation	<BASE>.wordseg.madcat.xml <BASE>.wordseg.sys.madcat.xml
	Line Segmentation (primary)	• transcription • translation • word-level segmentation	<BASE>.lineseg.madcat.xml <BASE>.lineseg.sys.madcat.xml
Document Image Recognition	Word Segmentation	• transcription • translation	<BASE>.wordseg.madcat.xml <BASE>.wordseg.sys.madcat.xml
	Line Segmentation (primary)	• transcription • translation • word-level segmentation	<BASE>.lineseg.madcat.xml <BASE>.lineseg.sys.madcat.xml
Document Text Translation	N/A	• translation	<BASE>.textseg.madcat.xml <BASE>.textseg.sys.madcat.xml